Molecular Informatics

QSPR modeling of liquid-liquid equilibria in two-phase systems of water and ionic liquid.

--Manuscript Draft--

Manuscript Number:	minf.202000001R2
Article Type:	Full Paper
Corresponding Author:	Kyrylo Klimenko Universidade Nova de Lisboa Faculdade de Ciencias e Tecnologia Caparica, Setubal PORTUGAL
Corresponding Author E-Mail:	alhimikir@gmail.com
Other Authors:	João Miguel Inês
	Jose Esperanca
	Luís Paulo N. Rebelo
	João Aires de-Sousa
	Gonçalo V. S. M. Carrera
Abstract:	The increasing application of new ionic liquids (IL) creates the need of liquid-liquid equilibria data for both miscible and quasi-immiscible systems. In this study, equilibrium concentrations at different temperatures for ionic liquid+water two-phase systems were modeled using a Quantitative-Structure-Property Relationship (QSPR) method. Data on equilibrium concentrations were taken from the ILThermo Ionic Liquids database, curated and used to make models that predict the weight fraction of water in ionic liquid rich phase and ionic liquid in the aqueous phase as two separate properties. The major modeling challenge stems from the fact that each single IL is characterized by several data points, since equilibrium concentrations are temperature dependent. Thus, new approaches for the detection of potential data point outliers, testing set selection, and quality prediction have been developed. Training set comprised equilibrium concentration data for 67 and 68 ILs in case of water in IL and IL in water modelling, respectively. SiRMS, MOLMAPS, Rcdk and Chemaxon descriptors were used to build Random Forest models for both properties. Models were subjected to the Y-scrambling test for robustness assessment. The best models have also been validated using an external test set that is not part of the ILThermo database. A two-phase equilibrium diagram for one of the external test set IL is presented for better visualization of the results and potential derivation of tie lines.
Response to Reviewers:	Dear Editor, The manuscript was revised in the light of new comments from the Reviewer 1 and a point-by-point response to critical comments is given below. New changes in the manuscript are highlighted in yellow. Sincerely yours, Kyrylo Klimenko Reviewer #1: The authors clearly answer my remarks. However, they did not systematically corrected the manuscript thus the same remark from the previous revision are still needed and the manuscript needs again a major revision. So the authors are invited to modify the manuscript as they have been invited to do. Page 4: "data [...] that cannot be converted to weight fraction...". The authors fail to explain up to this point why molecular weight is the relevant property to model, while molar fraction seems more adequate since it does not depend on the molecular weight of the IL. The authors must add some explanation in the introduction, before the state-of-the-art paragraph. -The short version of the answer to this question was added to the manuscript as follows: “Weight fraction was chosen for modelling, since it was important to distinguish between data from water-rich and IL-rich phase. Weight fraction always allows to do that in the simplest manner, i. e. considering data point with the weight fraction value below 0.5 indicates that the compound is a solute and if it is higher than 0,5, then it is a

	solvent (compound-rich phase). This is not straightforward using mole fraction." Page 4: "Structures were compared using both SMILES and InChi". SMILES and InChi notation are sensitive to graph ambiguities such as tautomers (although it is less true for InChi). Deduplication using these techniques are not recommended but I assume it was doable on simple structures such IL. The authors should add a word of warning in the manuscript at this point. -The response to reviewer was put into the manuscript. It is clear from eq 5 that as long as the standard deviation of the SALI is not null, one can chose coefA arbitrarily to give to the cutoff the value he wishes. So actually, the authors are taking a threshold value on a whim and seems to paint it as rational. So, if the authors wish to give a better understanding of their work, they must update the manuscript. -coefA and coefB may be considered artificial in the same sense as confidence intervals can be: it is conventional to use 90, 95 and 99% confidence intervals, rather than 90.5, 96 and 98.1%, even though the latter are not forbidden. In the same way, we do not expect users to apply 2.6 or 0.9 for coefA or coefB, respectively. Moreover, changing the default (3 and 1) coefA and coefB values is possible, however, it is not recommended and the default values were used in this study. We did not tune coefficients in order to get pre-conceived threshold values – this would be a research malpractice. The following statement was added to eq. 5 description: The software's default values of coefficients were used in this study. Concerning this remark, the authors confirmed my observation: the y-scrambling results cannot be considered "considerably less predictive", although it is understandable that the situation of IL in water is more difficult. -The word "considerably" was removed from the "The results of Y-scrambling model validation show that models based on randomized data show considerably less predictive capacity" statement in the description of the scrambling results. The following statement was added: "The IL in water models are less robust than the water in IL ones."
Additional Information:	
Question	**Response**
Submitted solely to this journal?	Yes
Has there been a previous version?	No
Dedication	

Title: QSPR modeling of liquid-liquid equilibria in two-phase systems of water and ionic liquid.

Authors: Kyrylo Oleksandrovych Klimenko* [a], João Miguel Inês [a], José Manuel da Silva Simões Esperança [a], Luís Paulo Nieto Rebelo [a], João Aires-de-Sousa [a], Gonçalo Valente Da Silva Mariño Carrera [a]

[a] *LAQV and REQUIMTE, Departamento de Química, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal*

* Corresponding author

Abstract:

The increasing application of new ionic liquids (IL) creates the need of liquid-liquid equilibria data for both miscible and quasi-immiscible systems. In this study, equilibrium concentrations at different temperatures for ionic liquid+water two-phase systems were modeled using a Quantitative-Structure-Property Relationship (QSPR) method. Data on equilibrium concentrations were taken from the ILThermo Ionic Liquids database, curated and used to make models that predict the weight fraction of water in ionic liquid rich phase and ionic liquid in the aqueous phase as two separate properties. The major modeling challenge stems from the fact that each single IL is characterized by several data points, since equilibrium concentrations are temperature dependent. Thus, new approaches for the detection of potential data point outliers, testing set selection, and quality prediction have been developed. Training set comprised equilibrium concentration data for 67 and 68 ILs in case of water in IL and IL in water modelling, respectively. SiRMS, MOLMAPS, Rcdk and Chemaxon descriptors were used to build Random Forest models for both properties. Models were subjected to the Y-scrambling test for robustness assessment. The best models have also been validated using an external test set that is not part of the ILThermo database. A two-phase equilibrium diagram for one of the external test set IL is presented for better visualization of the results and potential derivation of tie lines.

Keywords: Ionic liquids, ILThermo, phase diagrams, outlier detection

Introduction

The interest in ionic liquids (ILs) has increased over the last two decades due to some of their properties, [1] such as low melting points, [2] high electronic conductivities, [3] negligible vapor pressures [4,5], high thermal stabilities [6] and a set of complex molecular interactions related to electric charges, polarity and electronic structure, resulting in specific interactions (especially Coulombic and apolar interactions, as well as, hydrogen bonding).[7] These facts turn many ILs into well-marked nanostructured fluids, [8] making them useful for carbon dioxide capture, [9] azeotrope breaking [10,11] and extraction of bioactive compounds. [12,13]

The use of aqueous biphasic systems (ABS), as a promising extraction and purification medium for water-soluble molecules, particularly for biomolecules, is associated with high water concentrations. [14] The efficiency of the extraction will be determined by the equilibrium concentrations of the system. However, before predicting phase equilibria with 3 components it is sensible to try modeling of simpler, 2-component system, since its equilibria is already dependent on many parameters. One of the most important thermodynamic parameters that influences liquid-liquid phase equilibrium is temperature. Thermoresponsive water/IL mixtures can be divided into two main categories: those with an Upper Critical Solution Temperature (UCST) and those with a Lower Critical Solution Temperature (LCST) behavior. The character and magnitude of the temperature influence on the phase equilibrium varies significantly from one ionic liquid to another, which makes prediction of the equilibrium concentrations of the IL/water systems at different temperatures worth exploring.

Quantitative Structure-Property Relationship (QSPR) modeling of equilibrium concentrations of IL in water [15] has been carried out previously. However, in that study the data set has been restrained to only four ILs with the data in the narrow temperature range of 288.15 - 318.15 °K and water solubility in the IL was not examined at all. One of the major challenges of this work is to model a one-to-many relationship between the chemical compound and property data points, where one chemical has several equilibrium concentration values depending on temperature of experiment, unlike conventional one-to-one QS(P)AR studies. In previous QSPR

studies of temperature-dependent IL properties, the impact of temperature on the modeling approach was either ignored [16,17]

or separate models were made for every particular temperature [18,19]. In modeling of compounds other than IL, QSPR is applied to predict temperature-independent parameters, which are used in equations with direct temperature impact [20]. One-to-many relationships creates additional issues for the modeling as assessment of prediction accuracy, applicability domain definition and test set selection. [21] The latter can be exceptionally difficult, as shown by [22,23] in their attempt to model properties of mixtures and facing challenges to select a robust method for test set selection. In [24], researchers tried to overcome the "many-to-many" relationship issue in test set selection that arises from modeling properties of mixtures by doing external 5-fold cross-validation. This approach had some success, however it is time-consuming and authors acknowledge that the results of such external validation were overoptimistic, relatively to additional independent external validation.

Another challenge in modeling one-to-many relationship can be the presence of activity cliffs, or data points looking like that, due to heterogeneous data source. QSPRs can be either continuous (“activity hills”) or discontinuous (“activity cliffs”) based on whether small changes in compounds’ structure lead to small or dramatic changes in activity or, in our case, physico-chemical property. Small changes in molecular structure will cause small effects in the presence of gently rolling hills, or continuous SARs. This is in contrast to discontinuous SARs, where small changes in structure have dramatic effects. [25,26] Activity cliff estimation is also dependent on the descriptor space, since different descriptors reflect molecular properties differently.

Another issue concerns data points for different temperatures of the same compound originated from several sources that have little concordance, thus deteriorating the trend of temperature influence. If this happens, the model will have difficulties in predicting the equilibrium concentrations at different temperatures with high accuracy. The goal of this study is the development of QSPR models capable of predicting equilibrium concentrations for two phase water+IL systems in a broad range of temperatures (Figure 1).

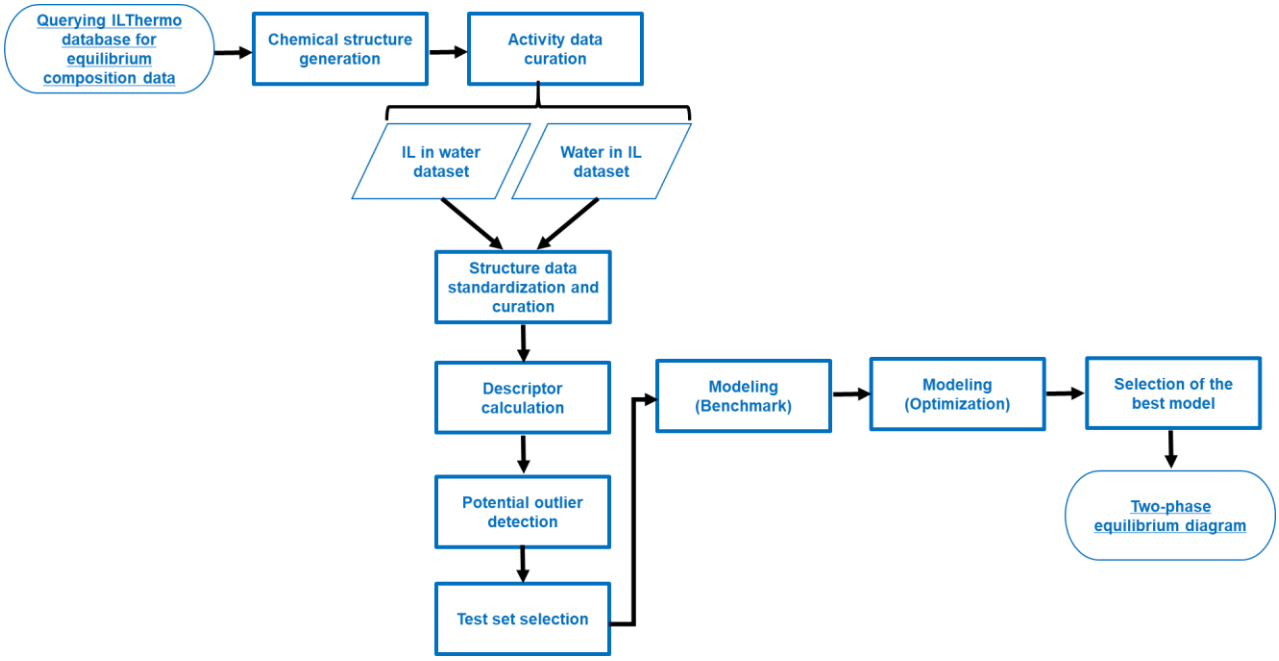


Figure 1 Workflow for the modeling of two-phase equilibrium for water+IL system at different temperatures.

Materials and Methods

Equilibrium concentration data was extracted from Ionic liquids database (ILThermo), [27] that consists of the published experimental data of thermodynamic and transport properties of ionic liquids, as well as, their binary and ternary mixtures. Data extraction was done using the pyILT2 package API. [28] The query is a function that makes use of a request module to carry out the search on the NIST server. The resulting JSON object is then decoded to a python dictionary. The reduced web form is composed of the mixture compound, number of components and the physical property of interest. By querying “H2O”, “binary mixture” and “composition at phase equilibrium” we have successfully collected a data set of 350 files.

For all ILs in the extracted data, both cations’ and anions’ IUPAC names were used as input to generate the SMILES representation using a Java library, OPSIN. [29] Only 13 out of 350 ionic liquid structures could not be parsed by this algorithm due to semantic issues occurring in the IUPAC string obtained from the ILThermo database due to semantic issues occurring in the IUPAC string obtained from the ILThermo database. Hence, a 96% coverage was accomplished by this method that turns to be a reliable technique for automatized data manipulation. The remaining 13 chemical structures

required manual inspection, through individual conversion of respective 2D structures into the SMILES file format.

Extracted data required curation both in terms of activity and structure. First, activity data were curated. Only data describing liquid-liquid equilibria with clearly determined temperature and pressure were kept. The decision was to model activity as weight (mass) fraction, which is a mass of constituent divided by the total mass of all constituents in the mixture [30]. Weight fraction was chosen for modelling, since it was important to distinguish between data from water-rich and IL-rich phase. Weight fraction always allows to do that in the simplest manner, i. e. considering data point with the weight fraction value below 0.5 indicates that the compound is a solute and if it is higher than 0,5, then it is a solvent (compound-rich phase). This is not straightforward using mole fraction. Thus, data in other units that cannot be converted into weight fraction were discarded: this applies to molality, molarity, mass per volume of solution and mass ratio to solvent. The data in molar fraction units was transformed into weight fraction. Then, the data was split into two data sets: 1) water in IL and 2) IL in water based on the value of the weight fraction of the solute: if the weigh fraction was higher than 0.5, then solute was re-labeled as solvent and data point transferred to the other data set after “1-weight fraction” transformation.

Structure standardization was done using Chemaxon Standardizer. [31] This includes standardizing the representation of aromaticity, mesomeric structure and functional groups. The full list of actions is given in Supplementary Material Table A1. Both data sets were examined for duplicate values, i. e. several data entries with the same structure and temperature value (group of duplicates). Structures were compared using both SMILES and InChi representations. SMILES strings and InChi notations came from standardized structure representation produced by Chemaxon Standardizer, where the same set of rules, including “Mesomerize” were applied to all compounds. The chances of overlooking a duplicate using SMILES and InChi under these conditions are very low. Conflicting weight fraction values in duplicates were looked upon carefully by inspecting the source publications. The definition of conflicting values is as follows: if absolute residual between standard deviation (SD) for the group of duplicates and mean SD for all duplicate groups is more or equal to SD of SD for all duplicate groups, then this group of duplicates has conflicting values. After conflicting values were examined, only one value was kept based on the

reliability of the source. Only one value from non-conflicting duplicates was kept as well.

The following molecular descriptors were generated in this study: 1) SiRMS [32] 2) MOLMAPS [33] 3) Chemaxon [34] 4) Rcdk. [35] SiRMS are fragment-based descriptors of varying length, atom and bond types and topology. In this study, fragments of size from 2 to 4 labeled by elements (atoms in fragments are either fully connected or 1 disconnection is allowed). MOLMAPs are descriptors based on physicochemical properties of fragments and were generated similar to Gupta et al. [33] but from atomic properties instead of bonds. In this study, 1) the descriptors were generated separately for the cation and anion 2) we used atom charge, orbital electronegativity, atomic polarizability , steric hindrance, H- bond donor and H-bond acceptor capacity as atomic properties 3) the Self-Organising Map (SOM) needed to generate the pattern of activated neurons was built using JATOON [36] software, based on 75% randomly selected atoms from the training set data 4) the SOM size was always 30x30 and the activation pattern was 1 0.75 0.5 0.25 for the central neuron first, second, and third level of neighbourhood, respectively. Both Chemaxon and Rcdk descriptors used in this study provide information on integral characteristics of the IL, such as molecular weight, lipophilicity, polar surface area, among other parameters. All Chemaxon and Rcdk descriptors are listed in Supplementary Material Table A4. In the end, four combinations of the above-mentoned descriptors (one fragment+one integral) were used as descriptor spaces for modelling: Chemaxon+SiRMS (ChSi), Rcdk+SiRMS (RcSi), Chemaxon+MOLMAPS (MoCh) and Rcdk+MOLMAPS (MoRc).

The distribution of the weight fraction values in both data sets is not normal, also poor solubility of most of the IL resulted in small property range with very small values, which can potentially complicate the modelling. In order to tackle that, dependent variables were normalized by cubic root transformation ($\sqrt[3]{\omega}$). Although distributions became less skewed, they were still abnormal, thus excluding the possibility to use the parametric statistic estimators (e.g. R^2, RMSE) for the model predictivity assessment. Two non-parametric error measures, namely Mean Average Error (MAE) and Mean Average Percentage Error (MAPE) were used in this study in two versions. The first consisted in treating all the data points as equal (eq. 1-2). The second, compound-based, consisted in processing all data points of a particular compound and then averaging the results from all compounds (eq. 3-4). This was done to tackle

potential bias against compounds with fewer data points having little impact on MAE/MAPE values, potentially leaving their misprediction unnoticed.

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \quad (1)$$

$$MAPE = \frac{100}{n} \sum_{i=1}^{n} \left| \frac{y_i - x_i}{x_i} \right| \quad (2)$$

$$MAE_{comp} = \frac{1}{k} \left(\sum_{j=1}^{k} \frac{\sum_{i=1}^{n} |y_{ji} - x_{ji}|}{n} \right) \quad (3)$$

$$MAPE_{comp} = \frac{1}{k} \left(\sum_{j=1}^{k} \frac{100}{n} \sum_{i=1}^{n} \left| \frac{y_{ji} - x_{ji}}{x_{ji}} \right| \right) \quad (4)$$

, where k is the number of compounds in the test set, n is the number of data points per compound, x_{ji} is the experimental result for the j-th compound at the i-th temperature. y_{ji} is the predicted result for the j-th compound at the i-th temperature.

Random Forest [37] (RF) was chosen as a machine-learning method and its implementation in R was used for the modelling. [38] RF is an algorithm consisting of a collection of tree-structured classifiers {h(x,), k=1, . . .} where the {} are independent identically distributed random vectors and each tree casts a unit vote for the most popular class (or in case of regression, average value) at input x. The out-of-bag (OOB) error estimation procedure was used for internal validation. The number of variables randomly sampled as candidates at each split (mtry) was optimized in every model in order to assure that the algorithm recognises the impact of the temperature on the equilibrium. There were 500 trees in every developed Random Forest model. It was decided to define the Applicability Domain (AD) based on the proximity between a training set data point and a predicted data point, i.e. how often a predicted data point ends up in the same terminal node of the tree as the particular training set data point. Distance between test set and training set data points have been used previously as a basis for AD [39], however in that case Euclidean distance was used to determine points proximity. Use of Euclidean distance was justified in that case, since model development was done with kNN algorithm that uses Euclidean distance to predict the property. However, the RF algorithm is based on different principles, thus in our case a RF-based proximity was used for AD definition. AD was defined as follows: if a data point has a terminal nodes proximity higher or equal to the threshold

value (0.5 for water in IL, 0.3 IL in water) with at least one training set data point, then the predicted data point is within AD. Threshold values differ due to difference in training set size between water in IL and IL in water equilibrium concentration data sets, in terms of the number of data points: the smaller size in case of IL in water makes it harder to find a training set data point analogue for a predicted data point.

Prior to test set selection, all data points were examined for activity cliffs formation potential. For every descriptor space, pairwise SAR analysis between all data point was done using the Structure-Activity Relationship Analyser (SARA), [**Error! Reference source not found.**, 41] in order to detect potential outliers. A data point must fulfil a criterion to be considered a potential outlier, i. e. it must be prone to form activity cliffs. The data point's potential to form activity cliffs, is determined as follows: first, $SALI^d$ (absolute difference between activities divided by Euclidean distance) was calculated using z-normalized descriptors for all possible pairs of data points in the data set. Next, a threshold value for $SALI^d$ was calculated according to eq. 5

$$Ac_{Thr} = \overline{SALI^d} + coefA \times \sigma_{SALI^d} \quad (5)$$

where, Ac_{Thr} is the threshold value for activity cliffs, $\overline{SALI^d}$ is the mean value of $SALI^d$s for all pairs in the data set, $coefA$ is a constant that determines how strict the threshold is, and σ_{SALI^d} is the standard deviation of $SALI^d$s for all pairs in the data set. If $SALI^d$ for two data points is higher than the threshold value, then the SAR between the two is considered to be an activity cliff. Then, occurrence of activity cliffs was calculated for every data point. A threshold for occurrence values was derived in the similar manner to the threshold of the eq. 5 – the mean occurrence value summed with the product of a coefficient with the standard deviation of the occurrence value. If the activity cliff occurrence value of a data point was higher than the occurrence threshold, then the data point was considered to be a potential outlier. The software's default values of coefficients were used in this study.

Potential to form an activity cliff, however, might not be enough to remove the data point from the data set, thus a manual inspection of data was carried out. Eight substances (1-methyl-3-octylimidazolium perfluorobutanesulfonate, 1-methyl-3-octylimidazolium tetrafluoroborate, 1-hexyl-3-methylimidazolium tetrafluoroborate, 1-decyl-3-methylimidazolium tetrafluoroborate, 1-hexyl-3-methylimidazolium tetracyanoborate, 1-ethyl-3-methylpyridinium bis--trifluoromethyl-sulfonylamide, 1-

ethyl-2-methylpyridinium_bis--trifluoromethyl-sulfonylamide, 1-butyl-3-methylpyridinium 1,1,1-trifluoro-N-[(trifluoromethyl)sulfonyl]methanesulfonamide) had data points marked as potential outliers within every descriptor space, so they were selected for further investigation. The plots that show weight fraction dependence on temperature are given in Supplementary Material Figure A1.

In three cases outlying data points were a result of high weight fraction values, one substance had a very steep increase in concentration with respect to temperature and one had no apparent reasons for forming discontinuous SARs. Data points from these substances were kept in the data set, since above-mentioned explanations do not provide enough ground for their removal. In three other cases analysis revealed one- or two data points that had weight fraction values that deteriorated the visible trend of equilibrium concentration dependency on temperature. Further investigation has shown that the outlying data point and the rest of the data point for a particular substance were produced by different research teams. Therefore, outliers can be a consequence of poor interlaboratory reproducibility. It was also discovered that two homologues of these substances (1-ethyl-4-methylpyridinium bis((trifluoromethyl)sulfonyl)amide, 1-ethylpyridinium bis[(trifluoromethyl)sulfonyl]imide) had outlying data point according to two out of four descriptor spaces, closer look at them showed that those substances have outliers. In the IL in water data set, data points from five ILs were marked as potential outliers (methylimidazolium tricyanomethane, 1-butyl-3-methylimidazolium hexafluorophosphate, 1-decyl-3-methylimidazolium tetrafluoroborate, 1-hexyl-3-methylimidazolium tetrafluoroborate, 1-methyl-3-octylimidazolium tetrafluoroborate). In four of them activity cliffs can be attributed to sharp increase in solubility at high temperatures, and, thus, data points were kept in the data set. One compound, however, had data points from different sources with no apparent concordance (Figure 2). It was decided to remove data points that correspond to this IL altogether.

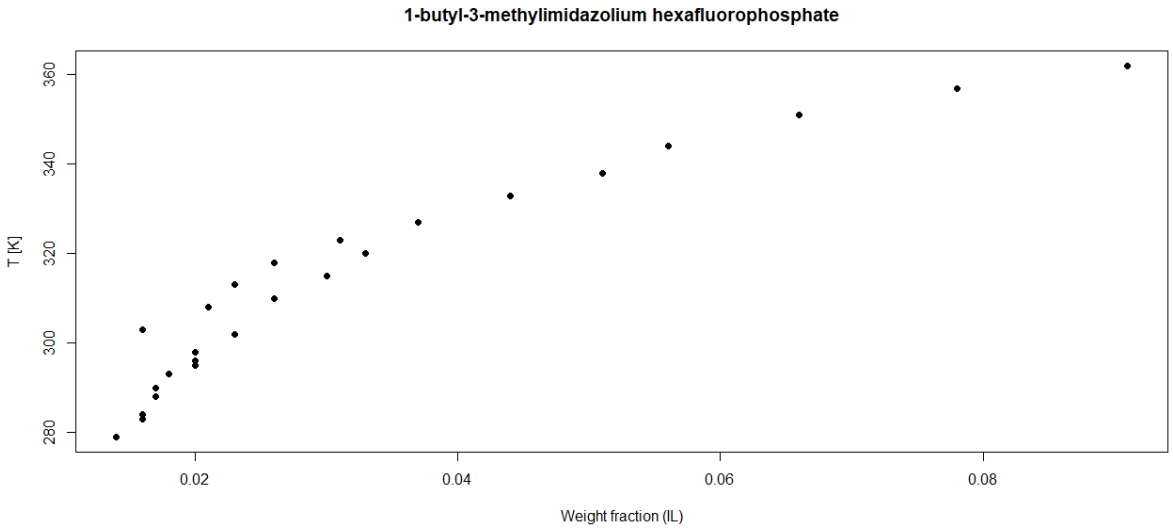


Figure 2 Weigh fraction (IL)-temperature plot for 1-butyl-3-methylimidazolium hexafluorophosphate resulting from the incorporation of data points from different sources with no apparent concordance.

Since there is a one-to-many relationship between substances and data point for this property, test set selection is crucial to the assessment of performance. Several test set selection schemes were tried with a Random Forest model trained with default parameters based on Chemaxon+SiRMS molecular descriptors. Other schemes were "random data point-out" and "random IL-out", which are modified approaches of published test set selection procedures for mixture properties, [22] where these methods were referred to as "points-out" and "mixtures-out", respectively. In 5-trials of 25% random data point-out selection, high reproducibility and accuracy were observed for the test set prediction (MAE range: from 0.0155 to 0.023, MAPE range: from 4.98% to 6.61%). However, this method has been criticized for overfitting the model since the test set, to a large extent, contains the same substances as the training set. 5-trials of 20% random IL-out have shown moderate accuracy and low reproducibility (MAE range: from 0.045 to 0.07, MAPE range: from 15.52% to 20.95%). It was decided not to pursue modeling with this scheme due to the lack of reproducibility, although IL-out seems to provide a more robust assessment, which is a desired feature. Thus, IL-out selection was modified by introducing selection criteria. Two schemes were designed:

a) IL-out based on the number of data points per IL

b) IL-out based on the number of data points per IL & structural similarity

In scheme a), the number of data points per every compound on the data set was calculated. Then, quartile values for number of data points distribution within the data set were calculated. Finally, 20% of compounds from each data subset formed by quartiles were selected for the test set randomly. Five modeling trials using this scheme showed almost no reproducibility (MAE range: from 0.025 to 0.082, MAPE range: from 7.76% to 28.63%). This approach was discarded.

The scheme b) was designed with the goal of selecting a subset that not only covers compounds with representative numbers of data points but also have similarities between ILs in the test and training set. In scheme b), after quartile values for the number of data points distribution within the data set were obtained, the Euclidean distance between substances within every subset formed by quartiles in the afore-mentioned z-normalized descriptor space (minus the temperature) was calculated. Then, for every subset, for every compound, median distance was found. Next, the median distance of the median distances was found and used as a threshold (eq. 6). In the end, the first 10% of compounds that had median values higher than the threshold and the first 10% of compounds that had median values lower than the threshold were selected for the test set.

$$Thr = \underset{j=1}{\text{median}}\left(\underset{i=1}{\text{median}}\left(x_{ij}\right)\right) \quad (6)$$

, where *Thr* is a threshold value, x_{ij} is Euclidean distance between the i-th compound to the j-th compound in the data subset ($i = j$ = *number of compounds in the subset*).

Computations using the scheme b) has led to a model with good predictive capacity (MAE = 0.0523, MAPE = 15.29%). This scheme was selected for the modeling that followed. However, the fact that there can be only one way to select test set compounds under this scheme makes robustness of this approach questionable and overfit is possible. Thus, an external test set had to be used to assess model performance in new situations.

Results & Discussion

Structure standardization and data curation, resulted in 548 data point (85 compounds) for water in IL data set and 424 data point (87 compounds) for IL in water

data set. Curated datasets have property range [0.002 - 0.48] and standard deviation of 0.069 for water in IL and range [0 - 0.33], standard deviation of 0.042 for IL in water. Main chemotypes of cations and anions are given in Supplementary Material Table A2 & A3 The composition of sets for model development is given in Table 1.

System	Overall	Training	Test	External test
water in IL	548 (85)	422-433 (67)	110-121 (18)	118 (17)
IL in water	424 (87)	318-321 (68)	78-81 (18)	84 (16)

Table 1 Number of data points and compounds (in brackets) in the modeling sets. The overall set is the QSAR-ready one, without the outliers removed. Internal validation for the random forest used all data points from the training set, since the out-of-bag method was used. External test set is comprised of data not included in the ILThermo database

Eight QSAR models were developed in this study based on the descriptor spaces mentioned before: four for IL in water and the four for water in IL. The mtry value was optimized in terms of OOB internal validation results (results are given in Supplementary Material Figure A2). The results of both internal validation and IL-out external validations are given in Table 2 and Table 3 for water in IL and IL in water systems, respectively. Model coverage is defined as the ratio between the number of predicted data points within the AD and the overall number of predicted data points.

Model	# of descr.	MAE (OOB)	MAPE (OOB)	# of ts data point	MAE (ts)	MAPE (ts)	MAE (ts per comp)	MAPE (ts per comp)	Model coverage
ChSi	657	0.003	0.87	110	0.0492	14.01	0.0484	16.78	1.00
MoCh	1226	0.0032	0.88	120	0.1040	26.51	0.1061	25.97	0.98
MoRc	1245	0.0029	0.8	121	0.0321	10.78	0.0428	12.46	0.76
RcSi	676	0.0031	0.88	116	0.0906	28.14	0.0647	22.13	0.85

Table 2 OOB validation and similarity-based IL-out test set validation results for the prediction of weight fraction of water in IL. descr. – molecular descriptors, ts – test set, per comp – per compound. Models are referred to by the descriptor space used.

Model	# of descr.	MAE (OOB)	MAPE (OOB)	# of ts data point	MAE (ts)	MAPE (ts)	MAE (ts per comp)	MAPE (ts per comp)	Model coverage
ChSi	625	0.0068	None[a]	78	0.0914	33.91	0.0604	23.36	0.67
MoCh	1801	0.0050	None	81	0.0571	28.48	0.0583	23.65	0.54
MoRc	1822	0.0041	None	78	0.0534	13.61	0.0417	17.48	0.35
RcSi	642	0.0068	None	78	0.0993	34.22	0.0621	23.97	0.73

Table 3 OOB validation and similarity-based IL-out test set validation results for the weight fraction of IL in water. descr. – molecular descriptors, ts – test set, per comp – per compound. Models are referred to by the descriptor space used. [a]MAPE (OOB) cannot be computed since some experimental results have weight fraction values of 0.

OOB validation results are quite optimistic for all the models because they incorporate predictions for data points obtained with trees trained with data points of the same compounds. However, accuracy of test set predictions varies enormously between the models with no particular space being the best. Using AD improves the predictivity for both IL in water, than water in IL models (see Supplementary Material Table A5 & A6 for statistics with no AD applied). Moreover, IL in water models have, in general, worse coverage than the water in IL ones (Figure 3).

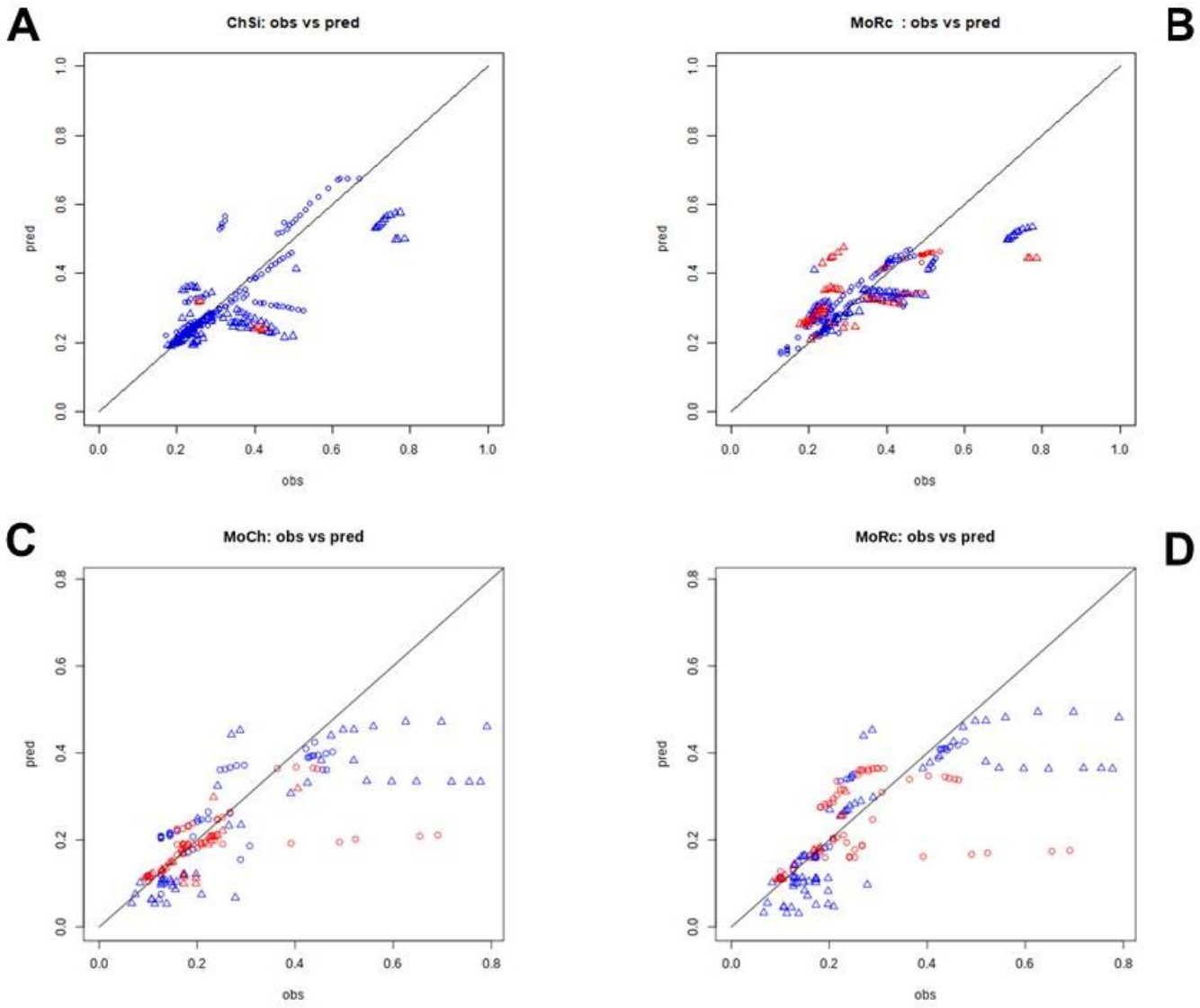


Figure 3 Observed vs predicted plots for the best models. A, B – water in IL models, C,D – IL in water models. Circles – test set, triangles – external test set, blue – within AD, red – out of AD.

The selection of the best models is the challenging due to the large number of statistical estimators that were used, all of which seem to highlight important aspects of models' performance. We decided that MAE and MAE per compound in combination with the model coverage must become the parameters the decision is based on. Thus, external test set was predicted by ChSi & MoRC models for water in IL systems and MoCh & MoRc for IL in water systems. The results are given in Table 4.

Model	# of ts data point	MAE (ts)	MAPE (ts)	MAE (ts per comp)	MAPE (ts per comp)	Model coverage
water in IL						
ChSi	118	0.0761	18.91	0.0819	19.64	0.95
MoRc	118	0.0730	18.95	0.0764	24.34	0.64
IL in water						
MoCh	84	0.1092	32.98	0.1094	35.45	0.57
MoRc	84	0.0849	30.99	0.0918	33.98	0.81

Table 4 External test set validation results for the prediction of the weight-fraction of water in IL and IL in water. Models correspond to describe spaces from Table 2 and Table 3. ts – test set, per comp. – per compound, data point – data points.

External test set results mostly show worse accuracy and coverage, and a closer look at prediction results revealed a potential problem with the proximity-based AD. When a new data point is predicted, it is characterized by the molecular descriptors and a temperature. The closest training set data point is also characterized by molecular descriptors and temperature. One could speculate that a certain data point that does not have one clear closest neighbor from the training set can still be predicted correctly, within the AD. Indeed, the training/test sets split was designed to have as much of the descriptor variety as possible. In many cases, data points are left out of coverage because of temperature, although similar structures were used in the training set, which enable reasonable predictions.

The skewness of the data set towards poorly miscible IL raised concerns about the models' ability to show chance correlation in prediction. Thus, a Y-scrambling, also known as y-randomization, [42] was done for ChSi (water in IL) and MoRc (IL in water). The dependent variable data was randomized 5 times in case of every data set, for both training and test set and modeling with RF was carried out with the same mtry optimization and statistical estimation as described above. The results are given in Table 5. Since re-modeling based on the scrambled data affects model coverage, the results are reported without AD restrictions and compared to the original modeling results without AD restrictions as well. The results of Y-scrambling model validation

show that models based on randomized data show less predictive capacity, then the original ones. The IL in water models are less robust than the water in IL ones.

	Y-R MoRc (IL)	MoRc (IL)	Y-R ChSi (water)	ChSi (water)
MAE	0.0919±0.0098	0.0749	0.1136±0.0027	0.0492
MAPE	None	24.54	34.95±3.53	14.01
MAE (per comp)	0.0869±0.0124	0.0534	0.1105±0.0079	0.0483
MAPE (per comp)	None	20.44	33.53±4.56	16.78

Table 5 Comparison of the Y-scrambling and normal models predictivity for the test set. Y-scrambling (Y-R) results are given with the standard deviation. MAPE for Y-R MoRc models could not be calculated since randomization have put some of the 0 values that were previously in the training set into the test set.

The skewness and range of both observed and predicted values raise another concern about the usefulness of models, particularly whether they can distinguish between highly soluble and insoluble IL, as well as, ILs high or poor capacity to dissolve water. Thus, a classification interpretation of the test set and external test set results was done for the four best models mentioned above. The data point was considered highly soluble if the weight fraction is equal or higher than 0.1. The classification statistics is given at Table 6. The results show that, in case of water in IL equilibria, ChSi model has good performance in both test set and external test set, whereas MoRc model performed well on the external test set but was unable to detemine highly soluble data points in the test set. None of the IL in water models performed well on the test set but MoRc showed acceptable results on the external test set. The problem with correctly classifying highly soluble compounds, especially in the case of IL in water, may be due to imbalanced data set, since highly soluble compounds are underrepresented.

	water in IL		IL in water	
	ChSi	MoRc	MoCh	MoRc

	ts	ext ts	ts	ext ts	ts	ext ts	ts	ext ts
Sensitivity	0.62	0.81	0	0.77	0	0.14	0	0.43
Specificity	0.93	1	0.98	1	1	1	1	1
Accuracy	0.86	0.97	0.89	0.96	0.95	0.78	0.96	0.88
BA	0.78	0.91	0.49	0.88	0.5	0.57	0.5	0.71
PPV	0.71	1	0	1	NA	1	NA	1
NPV	0.9	0.97	0.91	0.95	0.95	0.77	0.96	0.87
MCC	0.64	0.89	0.04	0.86	NA	0.33	NA	0.61
TP	15	13	0	10	0	2	0	6
TN	80	96	85	62	42	40	26	54
FP	6	0	2	0	0	0	0	0
FN	9	3	8	3	2	12	1	8

Table 6 Statistics on categorical interpretation of the results. ts is - set, ext ts - external test set. BA – Balanced Accuracy, PPV – Positive Predictive Value, NPV – Negative Predictive Value, MCC – Matthews Correlation Coefficient, TP – number of True Positive results, TN – number of True Negative results, FP – number of False Positive results, FN – number of False Negative results. Positive are values of highly soluble data points, Negative are poorly soluble ones. NA indicates a parameter that can not be computed.

Descriptor analysis based on variable importance in RF has shown the importance of temperature above all other descriptors. Molecular weight (MW) and logP (MlogP), number of rotatable bonds (Rotatable bond count, nRotB) are also present among the most influential descriptors. One out of two models have number of H-bond acceptors (acceptorcount) and topological polar surface area (TopoPSA) among the most important variables (Figure 4).

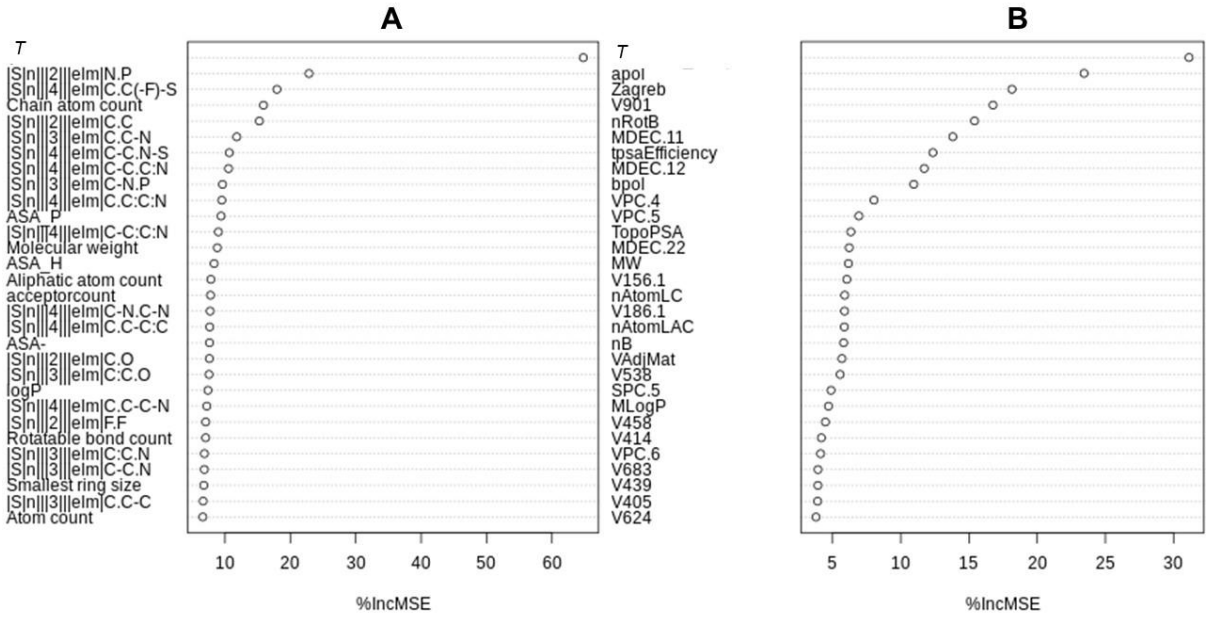


Figure 4. Variable importance of the best models. Importance is given as a percent of average increase in mean square error of the random forest as a result of variable values random shuffling. A – ChSi (water in IL), B – MoRc (IL in water).

Two-phase diagram for water + IL system

In order to have a better visualization of a water + IL system, a two-phase diagram was made for one of the external set compounds. Ethyl-dimethyl-propylammonium bis(trifluoromethylsulfonyl)imide was chosen as an example because it has data points in broad temperature range for both water in IL and IL in water phase (Figure 5). The diagram is made as follows:

- First, weight fractions for water in IL and IL in water are predicted for temperatures between 278-369 K – the range covered by ILThermo data sets. The conversion from the normalized values to original weight fraction is made and 1-weight fraction of IL in water transformation is applied to present all values in one plot.

- Next, if the weight fraction value is constant within a certain temperature range, then a median temperature value is used.

- After all predicted values are plotted, the experimental values are added to the graph as well, to show how close they are to the prediction.

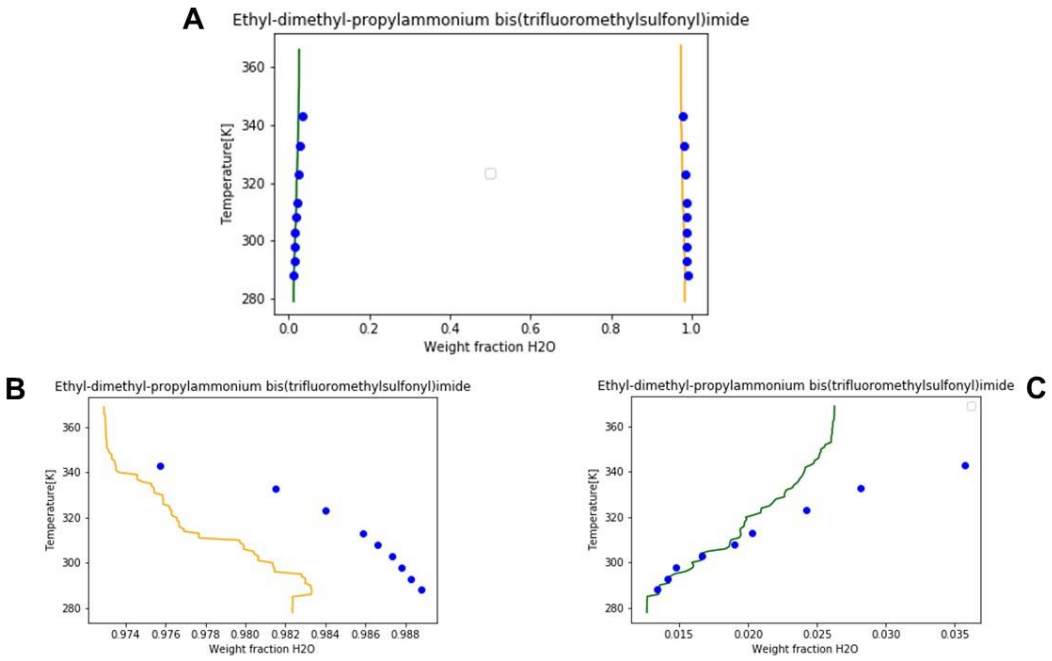


Figure 5 Example of a two-phase diagram generated using the best models of this study. A – the two-phase diagram over the whole weight fraction range, B – a zoomed plot for IL in water phase, C – a zoomed plot of water in IL. Yellow and green lines – predicted values, blue points – experimental values. More examples of two-phase diagrams are given for external test set ILs. The examples illustrate phase equilibria for ILs with experimental data points in both water-rich and IL-rich phases within a wide range of temperatures.

The figure shows that Ethyl-dimethyl-propylammonium bis(trifluoromethylsulfonyl)imide was predicted rather well. The fact that experimental values have a much more monotonous trend in temperature dependency can be attributed to the fact that they come from one source, whereas training set data was heterogeneous, sometimes with several local minima and maxima, which affected the ability of the model to predict temperature trends.

Ethyl-dimethyl-propylammonium bis(trifluoromethylsulfonyl)imide, just like most IL, is poorly miscible with water. This creates a problem for interpolation algorithms that are needed to close the envelop, since they will not be able to give a robust estimate for a point of convergence. However, the two-phase diagram representation is still useful for predicting tie lines of the particular system, since it provides equilibrium points in

both water-rich and IL-rich phases per every plausible temperature of liquid-liquid equilibrium of water+IL system.

This study, in line with the literature, shows that modeling a one-to-many relationships is a hard task, especially when data points come from different sources. An additional problem arises from the generally low mutual solubility of water and IL, that makes the distribution of equilibrium concentrations very narrow and skewed. Potential outlier detection and rational IL-out test set selection allow to overcome the problem to a certain extent, more efficiently for water in IL than IL in water models. This might be due to smaller number of data points in the IL in water data set.

Using a two-phase liquid-liquid equilibrium diagram for visualization provides experimentalists with an easy interpretation of predicted equilibria and allows to derive additional system parameters, such as tie lines. AD definition for these models are somewhat controversial and we encourage to trust the envelope prediction if at least one data point on both sides of the envelope are within the AD.

Acknowledgments
We thank Portuguese Foundation for Science and Technology Project: PTDC/EQU-EQU/30060/2017. This work was supported by the Associate Laboratory for Green Chemistry- LAQV which is financed by national funds from FCT/MCTES (UID/QUI/50006/2019) and co-financed by the ERDF under the PT2020 Partnership Agreement (POCI-01-0145-FEDER – 007265),

References

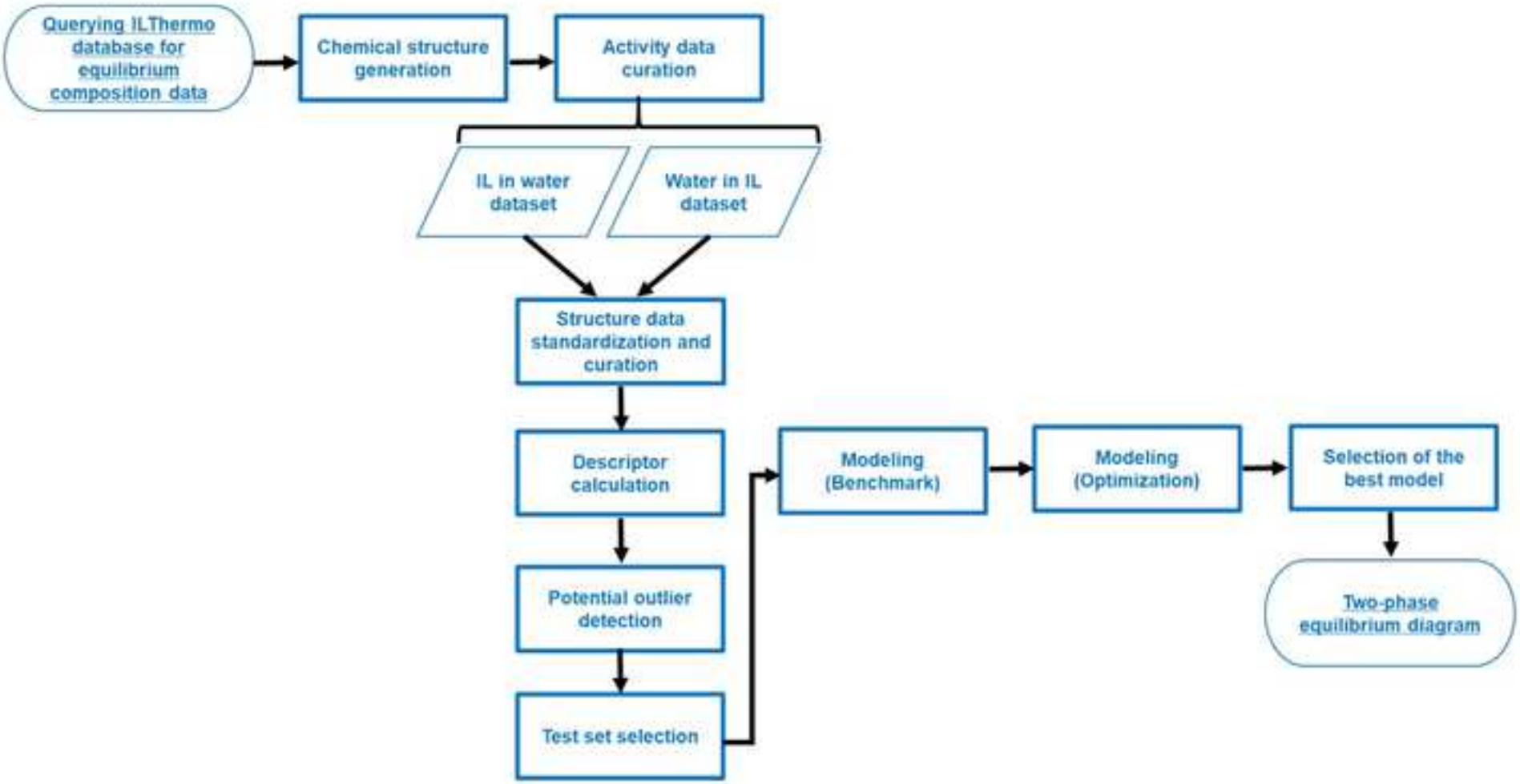
1. M. B. Shiflett, A. M. Scurto, *Ionic Liquids: Current State and Future Directions*, ACS Symposium Series, Vol. 1250, **2017**, Chapter 1 pp. 1-13 doi:10.1021/bk-2017-1250.ch001
2. P. M. Dean, J. M. Pringle, D. R. MacFarlane, *Phys. Chem. Chem. Phys.* **2010**, 12, 9144-9153 doi:10.1039/C003519J
3. O. Zech, A. Stoppa, R. Buchner, W. Kunz, *J. Chem. Eng. Data.* **2010**, 55(5), 1774-1778 doi:10.1021/je900793r
4. O. Aschenbrenner, S. Supasitmongkol, M. Taylor, P. Styring, *Green Chem*. **2009,** 11, 1217-1221 doi:10.1039/B904407H

5. M. J. Earle, J. M. Esperança, M. A. Gilea, J. N. Lopes, L. P. Rebelo, J. W. Magee, K. R. Seddon, *Nature*. **2006,** 439(7078), 831-834
6. C. Maton, N. *d.* Vosa, C. V. Stevens, *Chem. Soc. Rev.* **2013**, 42, 5963-5977 doi:10.1039/C3CS60071H
7. M. G. Freire, C. M. S. S. Neves, K. Shimizu, C. E. S. Bernardes, I. M. Marrucho, J. A. P. Coutinho, J. N. C. Lopes, L. P. Rebelo, *J. Phys. Chem. B.* **2010**, 114(48), 15925-15934 doi: 10.1021/jp1093788
8. A. B. Pereiro, M. J. Pastoriza-Gallego, K. Shimizu, I. M. Marrucho, J. N. Lopes, M. M. Piñeiro, L. P. Rebelo *J. Phys. Chem. B.* **2013** *117(37),*10826-10833*.*
9. Z. Zhao, H. Dong, X. Zhang, *Chinese J Chem Eng.* **2012**, 20(1), 120-129.
10. S. P. M. Ventura, F. A. e Silva, M. V. Quental, D. Mondal, M. G. Freire, J. A. P. Coutinho*, Chem. Rev.* **2017**, 117(10), 6984-7052 doi:10.1021/acs.chemrev.6b00550
11. K. S. Egorova, E. G. Gordeev, V. P. Ananikov*, Chem. Rev.* **2017**, 117(10), 7132-7189. doi: 10.1021/acs.chemrev.6b00562
12. F. S. Oliveira, R. Dohrn, A. B. Pereiro, J. M. M. Araújo, L. P. N. Rebelo, I. M. Marrucho, *Fluid Phase Equilibr*. **2016**, 419, 57-66. doi: 10.1016/j.fluid.2016.03.004
13. A. B. Pereiro, J. M. M. Araújo, J. M. S. S. Esperança, I. M. Marrucho, L. P. N. Rebelo, *J Chem Thermodyn.* **2012**, 46, 2-28 doi: 10.1016/j.jct.2011.05.026
14. B. Y. Zaslavsky, *Aqueous two-phase partitioning: physical chemistry and bioanalytical applications.* Vol. 99, M. Dekker, New York, **1995**, p. 694
15. M. G. Freire, C. M. S. S. Neves, S. P. M. Ventura, M. J. Pratas, I. M. Marrucho, J. Oliveira, J. A. P. Coutinho, A. M. Fernandes, *Fluid Phase Equilibr.* **2010,** 294(1-2), 234-240 doi: 10.1016/j.fluid.2009.12.035
16. A. Kondora, G. Járvása, J. Kontosa, A. Dallosa, *Chem. Eng. Res. Des.* **2014**, 92, 2867–2872
17. J. A. Lazzús, G. Pulgar-Villarroel *J. Mol. Liq.* **2015**, 211, 981–985
18. G. Yu, L. Wen, D. Zhao, C. Asumana, X. Chen, *J. Mol. Liq.* **2013**, 184, 51–59
19. C. Han, G. Yu, L. Wen, D. Zhao, C. Asumana, X. Chen, *Fluid Phase Equilibr.* **2011**, 300, 95–104
20. G. Cai, Z. Liu, L. Zhang, S. Zhao, C. Xu Energy Fuels, **2018**, 32, 3290−3298

21. K. Klimenko, V. Kuz'min, L. Ognichenko, L. Gorb, M. Shukla, N. Vinas, E. Perkins, P. Polishchuk, A. Artemenko, J. Leszczynski, *J Comput Chem*. **2016**, 37(22), 2045-2051 doi:10.1002/jcc.24424
22. I. Oprisiu, S. Novotarskyi, I. V. Tetko, *J Cheminformatics*. **2013**, 5, 1-4
23. I. Oprisiu, E. Varlamova, E. Muratov, A. Artemenko, G. Marcou, P. Polishchuk, V. Kuz'min, A. Varnek, *Mol. Inf*. **2012**, 31, 491 – 502
24. V. P. Solov'ev, I. Oprisiu, G. Marcou, A. Varnek, *Ind. Eng. Chem. Res.*, **2011**, 50 (24), 14162-14167
25. G. M. Maggiora, *J. Chem. Inf. Model*. **2006**, 46, 1535
26. H. Eckert, J. Bajorath, *Drug Discov Today*. **2007**, 12, 225-233.
27. A. Kazakov, J.W. Magee, R.D. Chirico, E. Paulechka, V. Diky, C.D. Muzny, K. Kroenlein, M. Frenkel, *NIST Standard Reference Database 147: NIST Ionic Liquids Database - (ILThermo), Version 2.0*, National Institute of Standards and Technology, Gaithersburg MD, 20899, http://ilthermo.boulder.nist.gov
28. Q. Dong, C.D. Muzny, A. Kazakov, V. Diky, J.W. Magee, J.A. Widegren, R.D. Chirico, K.N. Marsh, M. Frenkel, *J. Chem. Eng. Data*, **2007**, 52(4), 1151-1159
29. D. M. Lowe, P. T. Corbett, P. M. Rust, R. C. Glen, *J. Chem. Inf. Model.* **2011**, 51(3), 739-753
30. IUPAC. Compendium of Chemical Terminology, 2nd ed. (the "Gold Book"). Compiled by A. D. McNaught and A. Wilkinson. Blackwell Scientific Publications, Oxford (1997). Online version (2019-) created by S. J. Chalk. ISBN 0-9678550-9-8. https://doi.org/10.1351/goldbook
31. ChemAxon Standardizer v.19.2.0<https://www.chemaxon.com/products/standardizer/> (accessed 06.19).
32. V.E. Kuz'min, A.G. Artemenko, P.G. Polischuk, E.N. Muratov, A.I. Khromov, A.V. Liahovskiy, S.A. Andronati, S.Y. Makan, *J. Mol. Model*. **2005**, 11, 457-467.
33. S. Gupta, S. Mathew, P. M. Abreu, J. Aires-de-Sousa, *Bioorg. Med. Chem*. **2006**, 14(4), 1199-1206.
34. ChemAxon cxcalc plug-in v. 19.8.0 <https://chemaxon.com/marvin-archive/5_2_0/marvin/help/applications/calc.html> (accessed 06.19).
35. R. Guha, *J. Stat. Softw*. **2007**, 18, 1-16.
36. J. Aires-de-Sousa, *Chemometr. Intell. Lab*. **2002**, 61(1-2), 167-173. http://neural.dq.fct.unl.pt/jas/jatoon

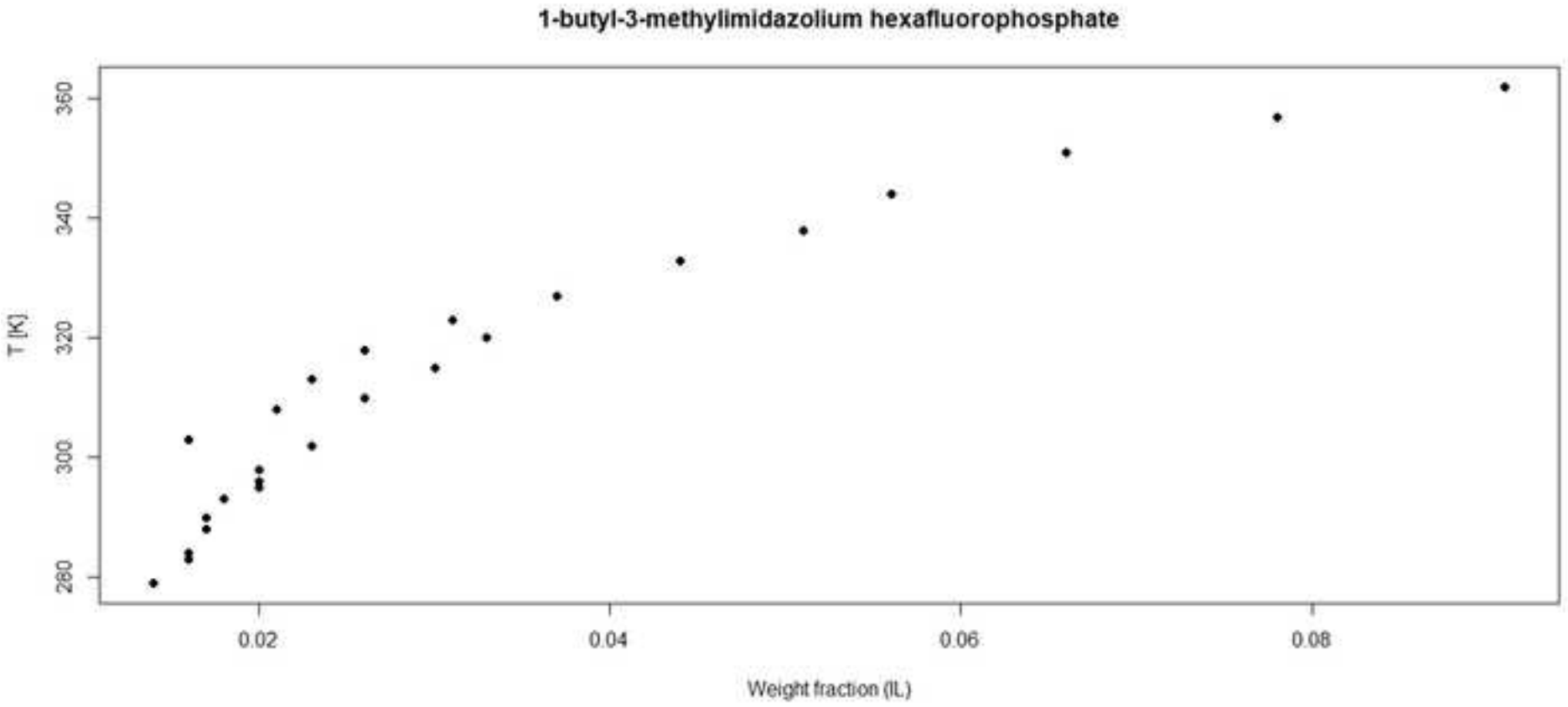
37. L. Breiman, *Mach. Learn*. **2001**, 45, 5–32..
38. A. Liaw, M. Wiener, *R News*, **2002**, 2(3), 18-22.
39. H. Zhu, A. Tropsha, D. Fourches, A. Varnek, E. Papa, P. Gramatica, T. Öberg, P. Dao, A. Cherkasov, I. V. Tetko, *J. Chem. Inf. Model.* **2008,** *48,* 766–784
40. K. Klimenko, *Mol. Inf.* ***2018****,* 37(4), 1-4.
41. Structure-Activity Relationship Analyser (SARA) v. 1.2, (https://github.com/klimenko-od91/SARA)
42. C. Rucker, G. Rucker, M. Meringer, *J. Chem. Inf. Model.* **2007,** *47,* 2345-2357

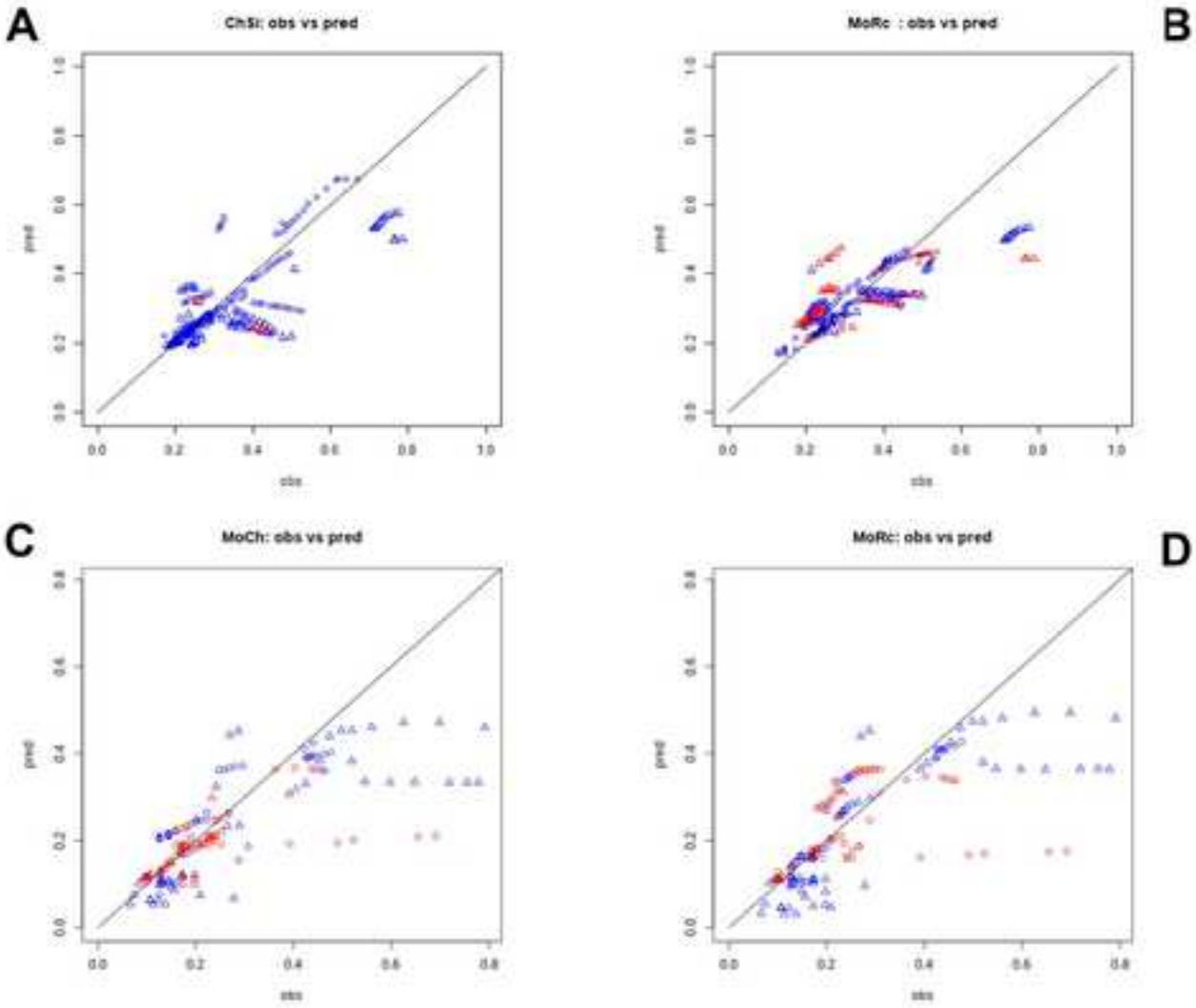
Graphical Material



Graphical Material

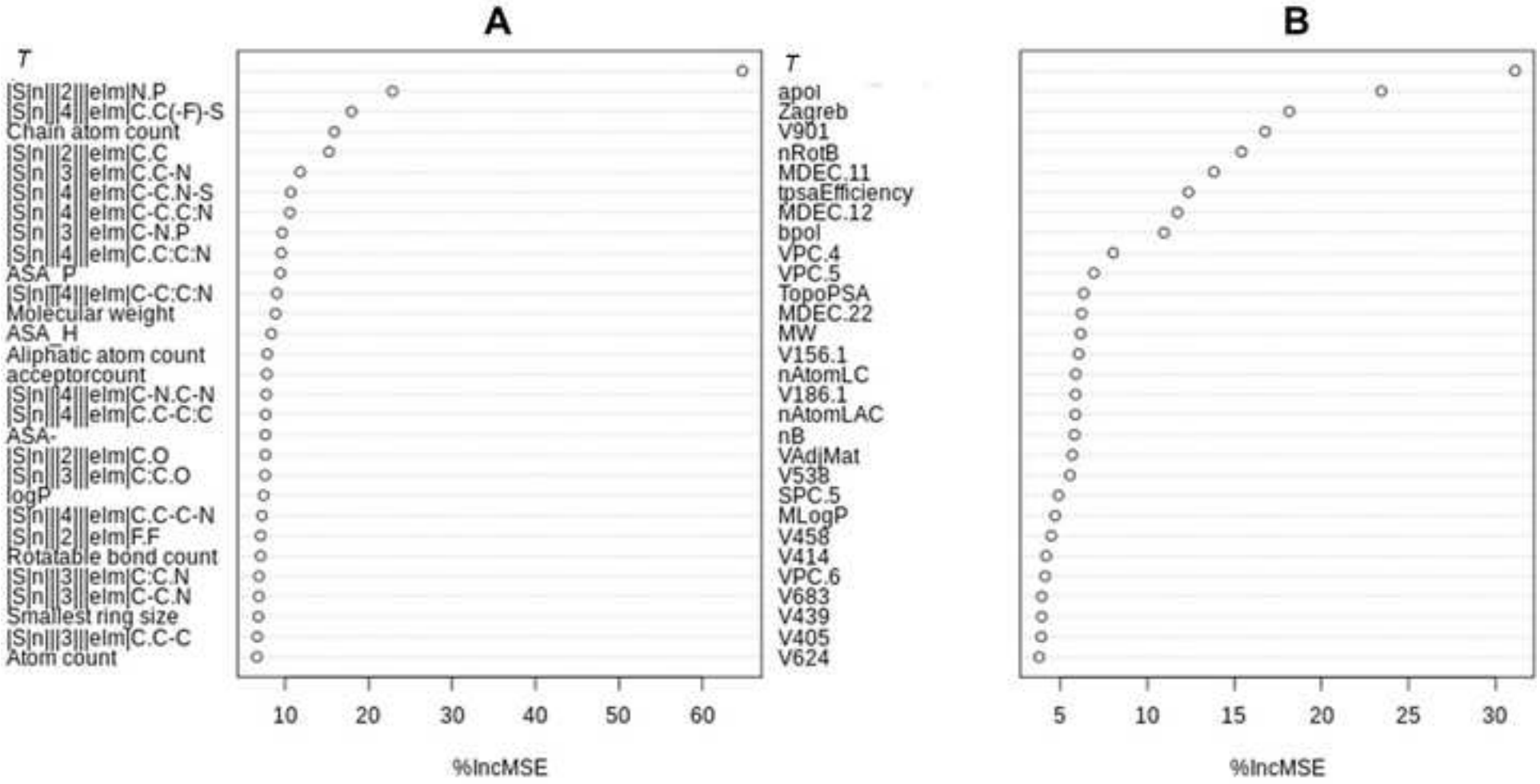
Click here to access/download;Graphical Material;Figure 2 newest.jpg





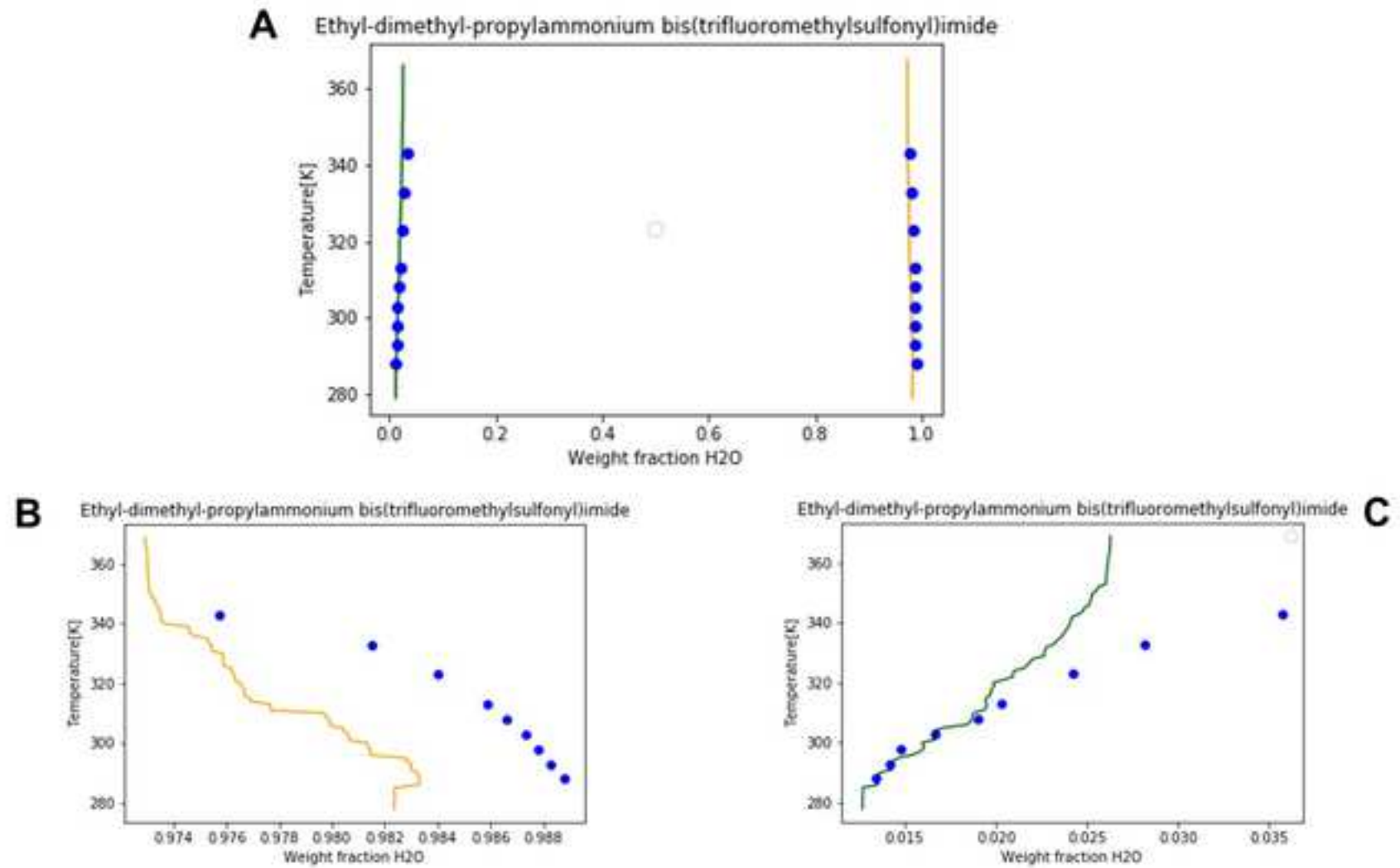
Graphical Material

Click here to access/download;Graphical Material;Figure 4.jpg



Graphical Material

Click here to access/download;Graphical Material;Figure 5.jpg



The source of external test set data (water in IL)

The source of external test set data (IL in water)